VISUALISATIONS AND NARRATIVES IN DIGITAL MEDIA

Methods and current trends

EDITORS

Pere Freixa, Lluís Codina,
Mario Pérez-Montoro & Javier Guallar

DigiDoc
Digital documentation and interactive communication Research group

Profesional de la información

Funding: This work is part of the Project INDOCS "Interactive storytelling and digital visibility in interactive documentary and structured journalism". RTI2018-095714-B-C21 (MICINN/FEDER). Spanish Ministry of Science, Innovation and Universities (Spain).

Contact:

DigiDoc Research Group
Communication Department
Universitat Pompeu Fabra
Roc Boronat, 138, 08018 Barcelona
digidoc@upf.edu
https://www.upf.edu/digidoc

Ediciones Profesionales de la Información SL
Mistral, 36 (Cortijo del Aire)
18220 Albolote (Granada), España
Tels.: +34-639 878 489 / 608 491 521
http://www.profesionaldelainformacion.com

Translations: Iain K. Robinson

1st edition, March 2022.
ISBN: 978-84-120239-9-2
DOI: https://doi.org/10.3145/indocs.2022.0
URL: http://www.profesionaldelainformacion.com/informes_digidoc_epi/visualisations_narratives_digital_media.html

Recommended citation:

Freixa, Pere; Codina, Lluís; Pérez-Montoro, Mario; Guallar, Javier (Ed.) (2022). *Visualisations and narratives in digital media. Methods and current trends*. Barcelona: DigiDoc Research Group & Ediciones Profesionales de la Información SL. https://doi.org/10.3145/indocs.2022.0

CONTENTS

Raquel Escandell-Poveda
Universitat d'Alacant
r.escandell@ua.es

Mar Iglesias-García
Universitat d'Alacant
mar.iglesias@ua.es

Natalia Papí-Gálvez
Universitat d'Alacant
natalia.p@ua.es

From Memex to Google: The origin and evolution of search engines

https://doi.org/10.3145/indocs.2022.4

Escandell-Poveda, Raquel; Iglesias-García, Mar; Papí-Gálvez, Natalia (2022). From Memex to Google: The origin and evolution of search engines. In: Freixa, Pere; Codina, Lluís; Pérez-Montoro, Mario y Guallar, Javier (ed.). *Visualisations and narratives in digital media. Methods and current trends*, (pp. 47-66). Barcelona: DigiDoc-EPI. https://doi.org/10.3145/indocs.2022.4

Abstract

Search engines have become one of the main channels for accessing information on the Web. Their widespread use means that the media, companies, institutions or any agent whose objective is to be visible to or to attract digital audiences is obliged to ensure its website is ranked at the top of the results pages. In Spain alone, 88% of people use search engines daily, with Google being the most used to conduct these searches, claiming a national market share of 95% and more than 90% worldwide. Given their current importance, an in-depth understanding of these tools becomes essential and it is worth pondering just how we have reached this current situation. What exactly are the origins of search engines and how has Google come to exercise its quasi-monopoly? The objective of this study is to explore the origin of search engines and to describe the main milestones in their evolution. To do so, a bibliographic review has been carried out using the main academic databases in the social sciences comprising a narrative search, complemented by an examination of the grey literature. The result is a journey through the history of search engines from their origins and subsequent technology developments to the creation of the World Wide Web. Likewise, a study is made of the original search engines and their main characteristics, with a particular emphasis on the path taken by Google given its current position of supremacy. Understanding the past has direct implications not only for our understanding of the present reality of web searches and access to information, but it is also essential for managing the continuous digital transformation to which we are all exposed, which has repercussions for all areas of the economy and society and, of course, for communication.

Keywords

Search engines, directories, Google, algorithms, SEO, web positioning, internet history.

Título

De Memex a Google. Origen y evolución de los buscadores

Resumen

Los buscadores se han convertido en una de las principales vías de acceso a la información existente en la Web. Su uso generalizado supone que medios de comunicación, empresas, instituciones o cualquier agente que tenga como objetivo la visibilidad o atracción de audiencias digitales, esté obligado a conseguir las primeras posiciones en los resultados de búsqueda. Solo en España, un 88% de personas utilizan buscadores diariamente, siendo Google el más utilizado, con una cuota en el mercado nacional del 95% y más del 90% mundial. Ante esta tesitura, se hace imprescindible un conocimiento profundo de estas herramientas y cabe preguntarse ¿cómo se ha llegado a esta situación? ¿Cuál es el origen de estos motores de búsqueda y cómo ha llegado Google a ejercer este cuasi monopolio? El objetivo de este estudio es explorar el origen de los buscadores y exponer cuáles han sido los principales hitos en su evolución. Para ello se ha realizado una revisión bibliográfica empleando las bases de datos académicas más destacadas en ciencias sociales así como una búsqueda narrativa, complementada con la consulta de literatura gris. El resultado es un recorrido por la historia de los buscadores desde su origen, el desarrollo de la tecnología y el surgimiento de la World Wide Web. Asimismo, se presentan los buscadores primigenios y sus características más destacadas, con especial énfasis en la trayectoria de Google por su posición de supremacía. Conocer el pasado tiene implicaciones directas no solo en el entendimiento del presente de las búsquedas y el acceso a la información, sino que es imprescindible para hacer frente a la transformación

digital contínua a la que estamos expuestos, la cual tiene repercusiones en todos los ámbitos de la economía, la sociedad y, por supuesto, en la comunicación.

Palabras clave

Buscadores, motores de búsqueda, directorios, Google, algoritmos, SEO, posicionamiento web, historia de Internet.

1. Introduction

The development of digital documentation, the emergence of the internet and the subsequent exponential growth of websites in the era of the World Wide Web have gone hand in hand with an intrinsic need for systems that can facilitate the management, filtering and retrieval of this information (Baeza-Yates and Ribeiro-Neto, 1999). Indeed, it was in response to this need that search tools originated and have since evolved in tandem with an ever-expanding Web. Today, as we enter the third decade of the 21st century, the Google search engine, with a world market share that climbs above 90% (Statista, 2021b), is responsible for 66% of web traffic in the USA (Fishkin, 2019), while, every second, 91,171 searches are conducted worldwide (Internet Live Stats, 2021).

In Spain, according to the *Asociación para la Investigación de Medios de Comunicación* (AIMC, 2020), internet penetration has reached 82.8% of the population and 88% of those use a search engine daily (Fernández, 2021). At the same time, search engines lead investment in digital advertising to the tune of 977.3 million euros, equivalent to roughly a third of the total (Interactive Advertising Bureau [IAB] Spain and PricewaterhouseCoopers, 2021), highlighting the importance of this channel for brands in the digital ecosystem.

The widespread use of search engines puts them in the crosshairs of any entity that needs to gain visibility and attract audience to its website (Escandell-Poveda, 2021). Based on these premises, this study seeks to explore the origins of search tools and learn about their history, by revisiting the many milestones in their evolution down to the present day.

The methodology we adopt in so doing is based on a review of the academic literature dedicated to our object of study, using for this purpose the Scopus and Web of Science databases, complemented by that of Google Scholar. More specifically, the review has been undertaken by conducting a narrative search and by examining the grey literature, while defunct web pages have been explored using the free-access Internet Archive Wayback Machine. The

results are discussed below under different headings, beginning with the origins of hypertext and concluding with a look at current figures for search engine use around the world.

2. Memex, the germ of the concept of hypertext

Humanity has had the need to organise information and to facilitate access to it since the founding of the first libraries. Around 1945, this need became even more pressing, as the military research undertaken during successive world wars boosted the development of scientific knowledge (García-Marco, 1996). Against this backdrop, Vannevar Bush (1945) envisioned a device he called the *memex* (Memory - Index), which would allow the storage and retrieval of information not included in the classic indexes. His system based the classification of documents on their links with each other and these same trails of reference would allow the user to jump from one document to another according to their relationship. Although the device never materialized, Bush's conception of the interrelationship of documents is considered the predecessor of *hypertext*, a term coined by Ted Nelson in the 1960s, and defined as an information management system in which data are stored in a network of nodes connected by links (Smith and Weiss, 1988).

In 1965, Nelson devised Project Xanadu, a global information storage system based on an electronic repository of interlinked documents offering immediate and simultaneous access to millions of people (Rayward, 1994), a futuristic version of the *World Wide Web* that would appear thanks to the efforts of Tim Berners-Lee in the early 90s.

3. The origins: the earliest search engines

Officially, the first *search engine* in history was Archie, created in 1990 by Alan Emtage, Bill Heelan and Peter Deutsch, students at McGill University in Montreal. Archie (a name that derives from the word *archive*) allowed queries to be made in its database comprising the names of files obtained from existing FTP servers. Archie did not index the contents inside the files, which meant searches were performed only on the file names (Schwartz et al., 1992; Seymour et al., 2011) (Image 1).

In 1991, as an alternative to Archie, Mark McCahill and his team created Gopher which permitted category searches to be performed and full texts accessed (Anklesaria et al., 1993). Gopher, considered the predecessor of future directories like Yahoo, displayed information in a structured fashion and allowed navigation between documents using links (Salazar, 2005). The problem was that it was limited to conducting searches in just one server, a drawback that was overcome with the development of Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) in 1992. This search engine provided results from the different

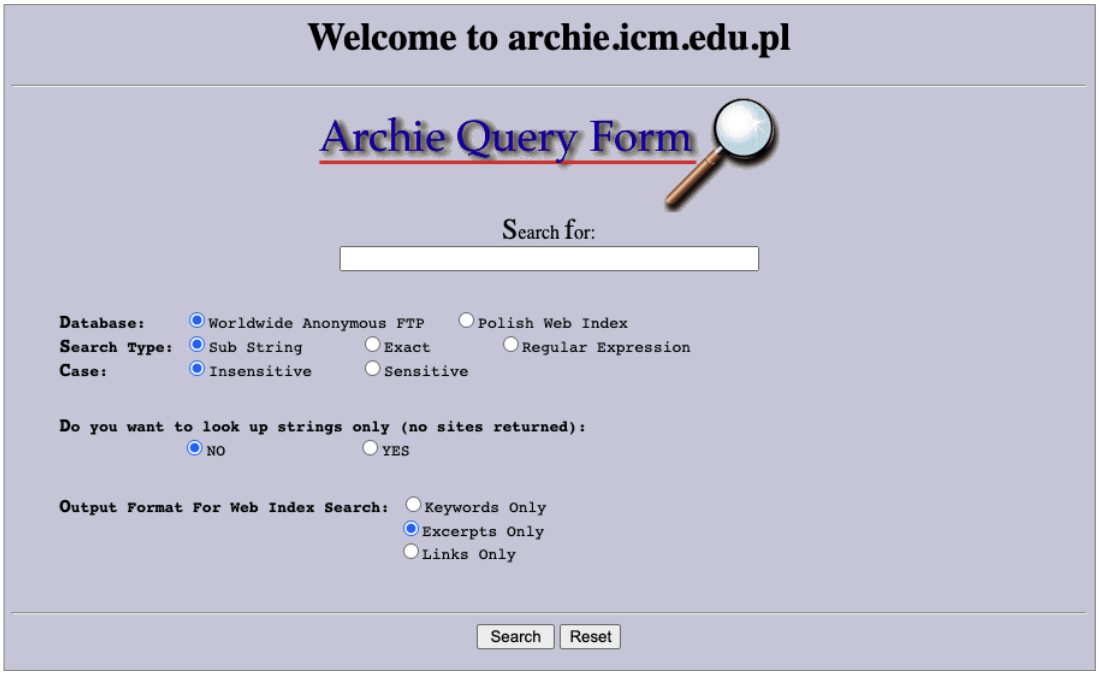

Image 1. *Archie's interface*. Source: http://archie.icm.edu.pl/archie-adv_eng.html.

Gopher servers by conducting searches with Boolean operators and keywords, although only on the menu items, without reaching the full texts. Jughead (Jonzy's Universal Gopher Hierarchy Excavation and Display), created in 1993, was another of the search engines used in "Gopher-space", although it differed from Veronica because it searched a single server at a time (Wall, 2017; Seymour et al., 2011; Smith and Updegrove, 1993; Sonnenreich, 1997; Wagner, 1996).

In parallel with these developments, Tim Berners-Lee, a scientist at CERN (European Council for Nuclear Research) was creating the *World Wide Web*, presented in 1989 as a hypertext-based system (Berners-Lee, 1989). It was created with the aim of facilitating access to information in Berner-Lee's own research, and became popular thanks to being open access, the development of HyperText Markup Language or HTML and the creation, in 1993, of the first graphical web browser called Mosaic (Griffiths, 2002; Tabarés-Gutiérrez, 2012).

With the arrival of the WWW, although lists of websites were available, such as the Index of content on the WWW Virtual Library, searches could not be conducted in them. To overcome this limitation, Oscar Nierstrasz created W3Catalog, the first World Wide Web search tool, which downloaded these resource listings and reformatted them to make them accessible through a searchable catalogue (Nierstrasz, 1996).

In June 1993, at the Massachusetts Institute of Technology (MIT), Matthew Gray developed the World Wide Web Wanderer, the first automated web crawler, whose objective was to chart the growth of the Web and discover new websites (Gray, 1996). The Wanderer robot

only visited the pages and counted them; it did not index them or save them in a database (Sherman and Price, 2001). A later extension allowed the robot to read web addresses or URLs (Uniform Resource Locators) as it went along, which generated Wandex, an index in which searches could be conducted and what is today considered the origin of web search engines as we now know them (Casares, 2008; Duklan and Bahuguna, 2012; Lara, 2014; Sonnenreich, 1997).

In October 1993, Martijn Koster created Aliweb, the Archie equivalent for Hypertext Transfer Protocol or HTTP. The fundamental difference with Wandex was that it did not use any robot crawler to create its index, rather the websites had to be registered manually (Koster, 1994). Its operation solved the problems of overloading the bandwidth that the trackers caused when entering the webs. In fact, it was Koster who created the "Robots Exclusion Protocol", a standard that prevents crawlers from entering web content through the robots.txt file (Jha et al. 2014; Koster, 1994). However, the fact of having to manually upload the websites to the directory limited its growth, since webmasters did not know how to create the necessary files to do so (Sonnenreich, 1997).

In December 1993, three search engines appeared based on automated crawlers or spiders similar to Gray's Wanderer: JumpStation, the first to combine the three essential features in one web browser: crawling, indexing and searching (Seymour et al., 2011), World Wide Web Worm and RBSE (Repository-Based Software Engineering). Like their predecessor, they only read certain fields or web tags and, the first two, presented their results in index date order. In this regard, therefore, RBSE was a pioneer, since it used a ranking based on the relevance of the search term (Eichmann, 1994; Mauldin, 1997; Sonnenreich, 1997).

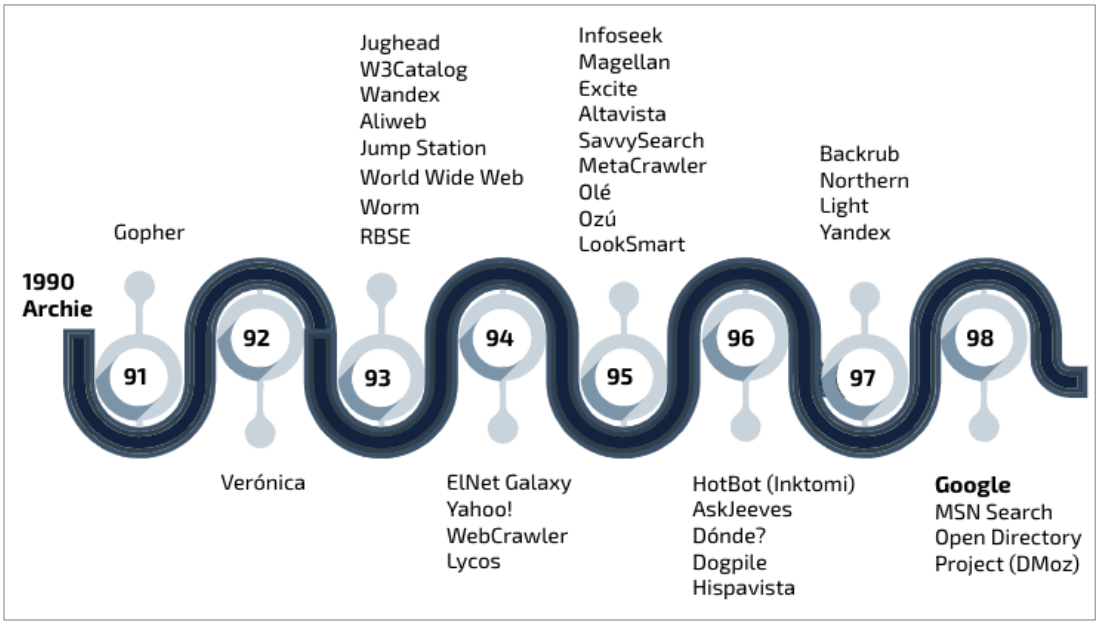

Image 2. Chronology of the earliest search engines.

The problem these early search engines faced was that the query had to use short concise words to ensure successful information retrieval. To overcome this limitation, in 1994 directories such as eiNet/Galaxy emerged, the latter considered the internet's first searchable directory (Galaxy, 2021). Web browsers of this type, rather than exploiting technological solutions, were able to create their database thanks to human editors who manually selected and classified the pages into categories and subcategories, thus allowing users to browse their index and explore the resources included on a specific topic (Sonnenreich, 1997).

Adopting this same philosophy, David Filo and Jerry Yang created Yahoo! in April 1994, a manually catalogued web directory that achieved great success and was one of the leaders in this market in the late 1990s (Trigo-Aranda, 2004). Yahoo! later evolved by incorporating a search engine that automated certain aspects of the collection and categorization of websites, thus blurring the distinction between search engine and directory (Casares 2008; Sonnenreich, 1997).

That same year, Brian Pinkerton created WebCrawler, a search engine that, unlike the then existing crawlers, indexed the entire content of the web (Seymour et al., 2011; Sonnenreich, 1997). In addition, it used a vector-based information retrieval model (Salton, 1989), which improved the results by displaying them according to relevance. This model took into account the frequency and weight of the query terms (Mauldin, 1997; Pinkerton, 1994), something that has since become the standard. Demand for WebCrawler was such that it crashed the servers of the University of Washington, where it had been created, and it was finally bought by the AOL portal (America Online), later being acquired by Excite (1997) and InfoSpace (2001). Lycos was another of the popular search engines launched in 1994. It stood out for its large number of indexed documents, which rose from 54,000 in July 1994 to 60 million in November 1996 (Mauldin, 1997) and its algorithm, which took into account the proximity between words and so improved the quality of its results (Sherman and Price, 2001). Lycos was also the first to display automatic excerpts on the search engine results pages (Sherman and Price, 2001).

In 1995 new competitors emerged in the market, among the most used were Magellan, Excite, Infoseek, Inktomi, Northern Light and AltaVista (Seymour et al., 2011). Infoseek became popular by becoming the default search engine on the Netscape navigator, while Excite introduced the concept of synonyms in its algorithm based on concepts, thanks to which results containing a related but not exact keyword could appear (Casares, 2008; Ballari, 2003).

AltaVista was the first to allow natural language querying, Boolean operators (AND, OR, NOT, ...) for queries, the possibility of searching all websites that linked to a specific page, the manual addition and deletion of websites, the conducting of searches in multimedia files and the inclusion of search tips, among other improvements (Seymour et al., 2011; Sonnenreich, 1997).

In 1995, with the curtailment of US government subsidies, privatization of the Internet began, quickly followed by massive investment in the *dotcom* commercial domains and with it the primary goal of the Web changed from being a place to provide information to serving as a place to attract customers (Ballari, 2003).

Search engines began to increase the size of their indexes, so that a search could produce hundreds of results, many of them irrelevant. Similarly, different search engines would return different results. This problem was addressed with the creation in 1995 of metasearch engines, such as MetaCrawler or SavvySearch, which simultaneously displayed the results of several search robots (Ballari, 2003; Schwartz, 1998; Sonnenreich, 1997). Faced by this new situation, website owners began to manipulate the results so as to appear at the top of the rankings (Ballari, 2003), thus laying the foundations for SEO (Search Engine Optimization).

HotBot, launched in 1996, was noted for its speed and power, being considered the first search engine capable of indexing the millions of websites that then existed, as well as using cookies to return personalized results per user (Casares, 2008; Seymour et al., 2011; Sonnenreich, 1997). That same year saw the launch of Ask Jeeves, which answered questions posed in natural language, and BackRub, the search engine that would be renamed Google in 1997, was started by two students at Stanford University (Brin and Page, 1998). Google began to grow in popularity, becoming a world leader in the early 2000s and has since maintained its dominance over its competitors, acquiring a global market share of between 80 and 90% in the second decade of the 21st century (Johnson, 2021).

The year 1997 saw the launch of Northern Light, a search engine that introduced search suggestions, but which was subsequently developed for private custom only, and Yandex, which became the largest Russian search engine (Casares, 2008; Seymour et al., 2011) until it too was surpassed by Google in 2017 (Liveinternet, 2021).

4. Search engines of the new century

In the late 1990s, search engines also formed part of the investment frenzy attributable to the internet boom. Like a multitude of companies at that time, many saw their finances plummet when the *dotcom* bubble burst between 1999 and 2001 (Seymour et al., 2011). In 1999 it was estimated that the public Web contained more than 800 million pages but that no single search engine was capable of indexing more than 16% of that total (Lawrence and Giles, 1999) and it was predicted that this number would soon reach thirteen billion pages (Sherman, 2000).

In 1999, AlltheWeb appeared, establishing itself as the owner of the largest Internet database, as did GoTo (later becoming Overture), the first to offer pay-per-click advertising (Fain and Pedersen, 2006; Van Couvering, 2008). In 2000, the Teoma search engine was launched, using

a system of link popularity, similar to Google's, based on "ExpertRank" technology. What made it different was it took into account the context of the link, giving it more importance if the topic of the page was related to that of the website receiving it (Casares, 2008; Seymour et al., 2011). In that same year, Baidu, the Chinese and Japanese website, image and audio search engine was also launched, and which in 2021 continues to lead the search engine market in China (StatCounter, 2021a).

Despite the success enjoyed by Google from the early years of that decade – further boosted by the fact that Yahoo! included Google as a search engine on its homepage (Baker, 2021) – new search engines continued to appear: In 2000, Exalead was created in France, displaying thumbnail previews and providing advanced refining of its results; in 2003, Seekport was launched, offering local, regional versions adapted to each country; in 2004, Nutch, an open source web crawler, was created; and, in the same year, Clusty, which grouped results into thematic categories (or clusters), and Yahoo! Search, the search engine evolution of the Yahoo! Directory, were released (Seymour et al., 2011).

In 2005, the Microsoft search engine, MSN, was also released. Its name would later be changed, first, to Windows Live Search in 2006 and then to Bing in 2009.

Other ventures – most notably GoodSearch (2005), Forestle (2008) and Ecosia (2009) – opted for a different business approach, donating part of their profits to charities and non-profit organizations. In 2006, Hakia, a pioneer semantic search engine, was launched. In this respect it differed from all of its competitors who continued to work by keywords, albeit that, since 2003, Google had begun to implement improvements in its semantic searches (Casares, 2008). Natural language processing, aimed at improving the quality of results, was

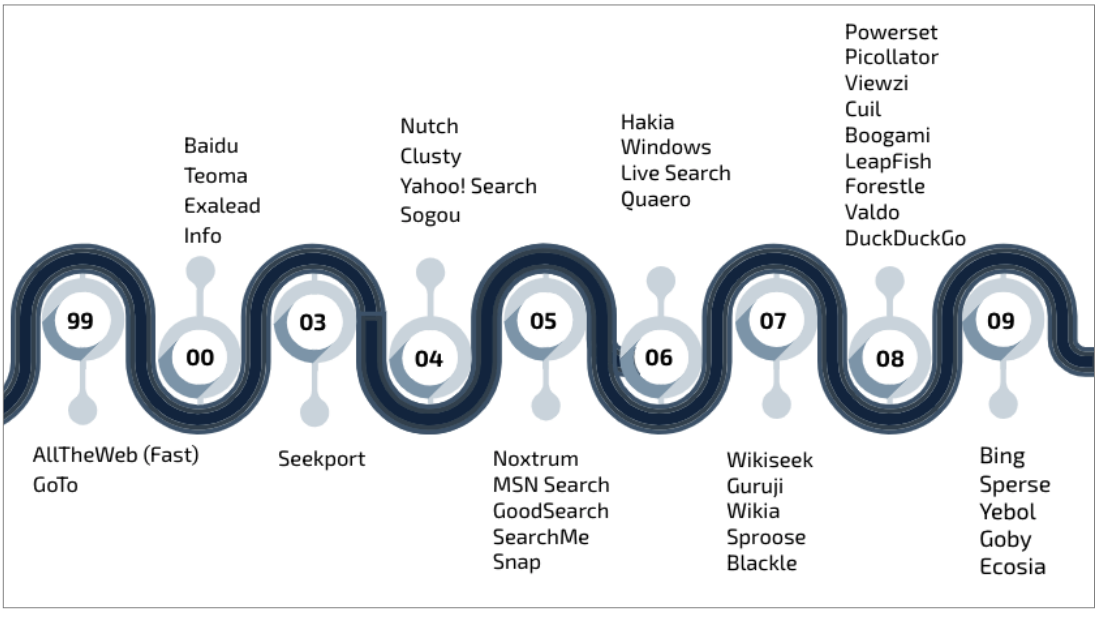

Image 3. Chronology of the main search engines launched between 1999 and 2010.

one of the developments made in those years (Pérez-Agüera, 2008), something that Google would continue to develop in its algorithm with the following updates: Brandy in 2003, Hummingbird in 2013 and Bert in 2019.

5. The arrival of Google

In 1995, on matriculating at Stanford University to pursue his PhD in the Computer Science Department, Larry Page met Sergey Brin and, the following year, both students began tracking and downloading websites and analysing their links (Vise and Malseed, 2006). Page worked on the theory that the more links a website received the more important it was, basing his thinking on the citation system used in the scientific literature, in which having more citations was equivalent to being more popular. However, he understood that not all links had the same weight, but that those originating from pages with more links were more important. He called this link classification system PageRank, a play on words based on his own surname, and used it in the creation in 1997 of a search engine which he called BackRub (Vise and Malseed, 2006; Wills, 2006).

That same year, Page, Brin, and Rajeev Motwani, the Stanford professor who tutored him, worked on the tool and built a prototype search engine that added PageRank to the technology then being used by search engines. All told, they disposed of a repository that stored 24 million pages and which implemented PageRank, building an index of links at the same time as the web crawler systematically browsed the sites (Page et al., 1999).

In the Autumn of 1997, they changed the name of their search tool to Google. On September 15, they registered the Google.com domain, and made it available to the community at Stanford University via the url: www.google.standford.edu (Vise and Malseed, 2006) (Image 4).

Their premise was that existing search robots basically relied on keyword matching, which returned too many low-quality results (Brin and Page, 1998). Given the exponential growth in the number of websites, for Brin and Page, it was critical to obtain more precise results, something they achieved by taking into account the link structure; the link (anchor) text, which worked as a description of the specific page that a link points to; and, the proximity information about the elements of a document (Brin and Page, 1998).

PageRank worked as an objective measure, assigning importance to a website according to the number and weight of each of the citations. To calculate a page's PageRank, they took into account the pages that link to it, the number of outgoing links from that page and the damping factor, which is normally set to 0.85, and which refers to the probability that on any given page a surfer continues to click on the links it contains or stops clicking and moves on to another page. Thus, a page will have a higher PageRank the more links it has or if, even if it

Image 4. Google interface in 1998. Source: https://web.archive.org/web/19981111183552/http://google.stanford.edu/

does not have many links, the few links it receives are from pages with a high PageRank (Brin and Page, 1998).

This combination of factors permitted the development of a system of ranking results by importance and not by a simple connection between the words of the query and those contained on the website, which meant the results were much more accurate (Vise and Malseed, 2006).

In March 1998, they tried to sell the PageRank patent to existing search engines, including AltaVista, Excite and Yahoo but all of them rejected the opportunity, preferring Internet users to stay in their portal rather than have them leave to visit other pages (Vise and Malseed , 2006). In August of that same year, they obtained the necessary financing and a month later they founded Google Inc (Google, n.d.; Wills, 2006).

Google was to revolutionise the world of search engines as it was then known (Seymour et al., 2011) and the company gradually gained market share until establishing itself as the undisputed leader. Indeed, in 2003 it already accounted for more than half of the searches made (Trigo-Aranda, 2004) and today, it has maintained its position, well above the rest of the search engines, with a worldwide use that reaches more than 86% (Johnson, 2021).

6. Evolution of the Google algorithm

Among the various modifications made to the operation of search engines, the launch of Google's PageRank algorithm in the 1990s was the one that revolutionised the industry.

The incorporation of this innovation made the manipulation of search results much more difficult, as they no longer depended on website creators, but rather on third parties who had to link to the sites' pages to increase their popularity and, hence, a page's positioning (Arbildi-Larreina, 2005; Brin and Page, 1998; Morato-Lara et al., 2013).

From the very outset, Google's algorithm has been subject to constant updates. According to Moz (n.d.), in 2018 alone, 3,234 updates were made. In 2003, the first named update – Boston – was released, involving the implementation of monthly structural changes in the algorithm combined with important updates of the index, a process that became known as the *Google Dance* as it introduced sizeable ranking shifts (Baker, 2021).

Among the most important updates (Image 5), we find, in chronological order, that of Florida, in 2003, which penalizes keyword *stuffing* and other abusive techniques carried out to manipulate positioning. Brandy was launched in 2004, which, among other modifications, introduced latent semantic indexing (LSI), which expanded Google's ability to understand synonyms and keyword context and began the rating of the relevance of anchor texts, the text containing a hyperlink (Moz, n.d.).

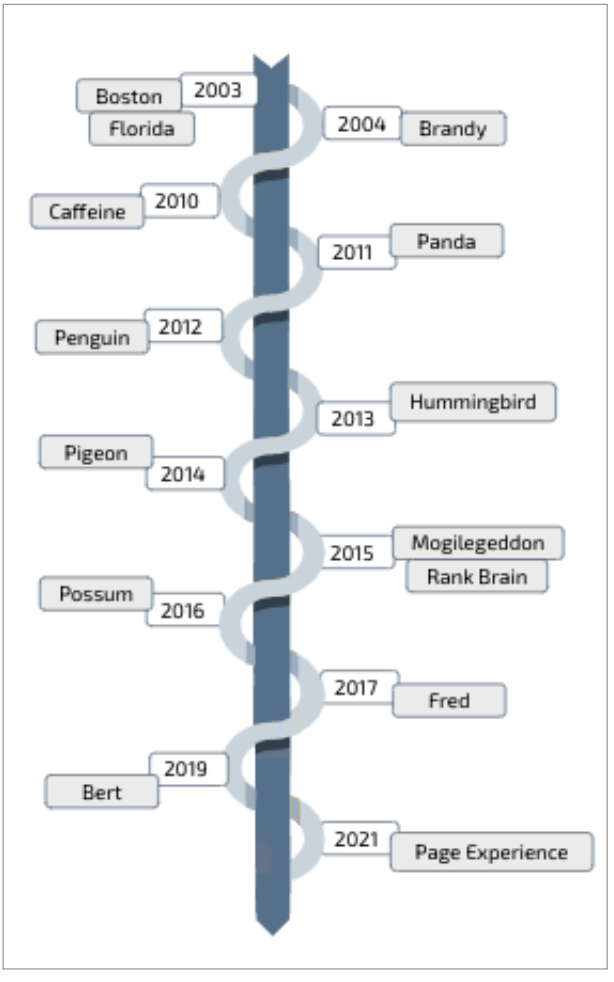

Image 5. Main updtaes to the Google algorithm.

In 2010, the Caffeine update (Bezhovski, 2015; Cutts, 2009; Gonzalo-Penela, 2015; Grimes, 2010; Morato-Lara et al., 2013; Shahzad et al., 2020) provided for a faster indexing of content while search results were now more current than before by up to 50 per cent. And a year later, in 2011, Panda was launched for the first time, one of the most influential updates since it penalizes websites with little or poor quality content, such as the so-called content farms, and those with duplicate and/or copied content (Aswani et al., 2018; Bezhovski, 2015; Egri and Bayrak, 2014; Giomelakis and Veglis, 2016; Gonzalo-Penela, 2015; Goodwin, 2018; Gudivada, 2015;

Mavridis and Symeonidis, 2015; Nanda et al., 2019; Shahzad et al., 2020; Singhal, 2011; Zhang and Cabage, 2017).

In 2012, another of the major updates was made, Penguin. This penalized a series of overoptimization practices considered webspam, such as keyword stuffing – thereby, renewing the 2003 update – and the buying and trading of links (Aswani et al., 2018; Bezhovski, 2015; Egri and Bayrak, 2014; Giomelakis and Veglis, 2016; Gonzalo-Penela, 2015; Gudivada, 2015; Nanda et al., 2019; Shahzad et al., 2020; Zhang and Cabage, 2017). In 2013, Hummingbird represented a complete update of the algorithm to provide more relevant results thanks to a better understanding of natural language and voice queries (Aswani et al., 2018; Bezhovski, 2015; Giomelakis and Veglis, 2016; Gonzalo-Penela, 2015; Goodwin, 2018; Gudivada, 2015; Lopezosa et al., 2018; Mavridis and Symeonidis, 2015; Nanda et al., 2019; Shahzad et al., 2020) and, in 2014, Pigeon, introduced drastic changes in how local searches are interpreted (Aswani et al., 2018; Gudivada, 2015; Nanda et al., 2019).

In 2015, the main updates went by the name of "Mobilegeddon", which involved a transformation in the rankings of mobile search results, giving priority to websites optimised for viewing on these devices (Goodwin, 2018; Shahzad et al., 2020), and Rank Brain, by which Google announced that machine learning now formed part of the algorithm, that is, artificial intelligence that permits the system, by means of automatic learning, to better understand the intentions of search queries (Lopezosa et al., 2018; Rovira et al ., 2018; Nanda et al., 2019; Shahzad et al., 2020). The following year, Rank Brain was confirmed by Google as one of its three most important ranking factors (Schwartz, 2016; Shahzad et al., 2020).

The Possum and Fred updates are assumed to have been released in 2016 and 2017, respectively, although they have never officially been confirmed by Google (Moz, n.d.). Possum gives relevance to local businesses, using a user's geolocation (Hawkins, 2016; Nanda et al., 2019), while Fred penalizes websites with excessive advertising or content created exclusively for monetization, that is, the generation of traffic income from advertisements or affiliate commissions (Gonzalez, 2017; Nanda et al., 2019).

In 2019, the BERT (Bidirectional Encoder Representations from Transformers) update was introduced, aimed at improving understanding of the language used in searches and giving better matches between queries and results (Devlin et al., 2019; Nayak, 2019).

In June of that same year, the Google Page Experience update was rolled out. This incorporates the so-called Core Web Vitals, which measure a website's speed and response factors so as to provide an optimal browsing experience (Google, 2021; Gupta, 2021; Schwartz, 2020).

7. Search engines around the world

The importance of search engines as a means of accessing internet content is clearly evident if we consider their penetration rate around the world. In countries like Belgium, Norway and South Africa they are used by 85% of the population (Figure 1), while in Spain the percentage coincides with that of the world mean, with 74% of Spaniards using search engines frequently (Statista, 2013).

Among the search engine options available worldwide, Google is the most used, with a market share of 91.4% (Statcounter, 2021b) (Figure 2). The degree of quasi monopoly enjoyed by Google varies from country to country. In Russia, its market share is not as pronounced due to

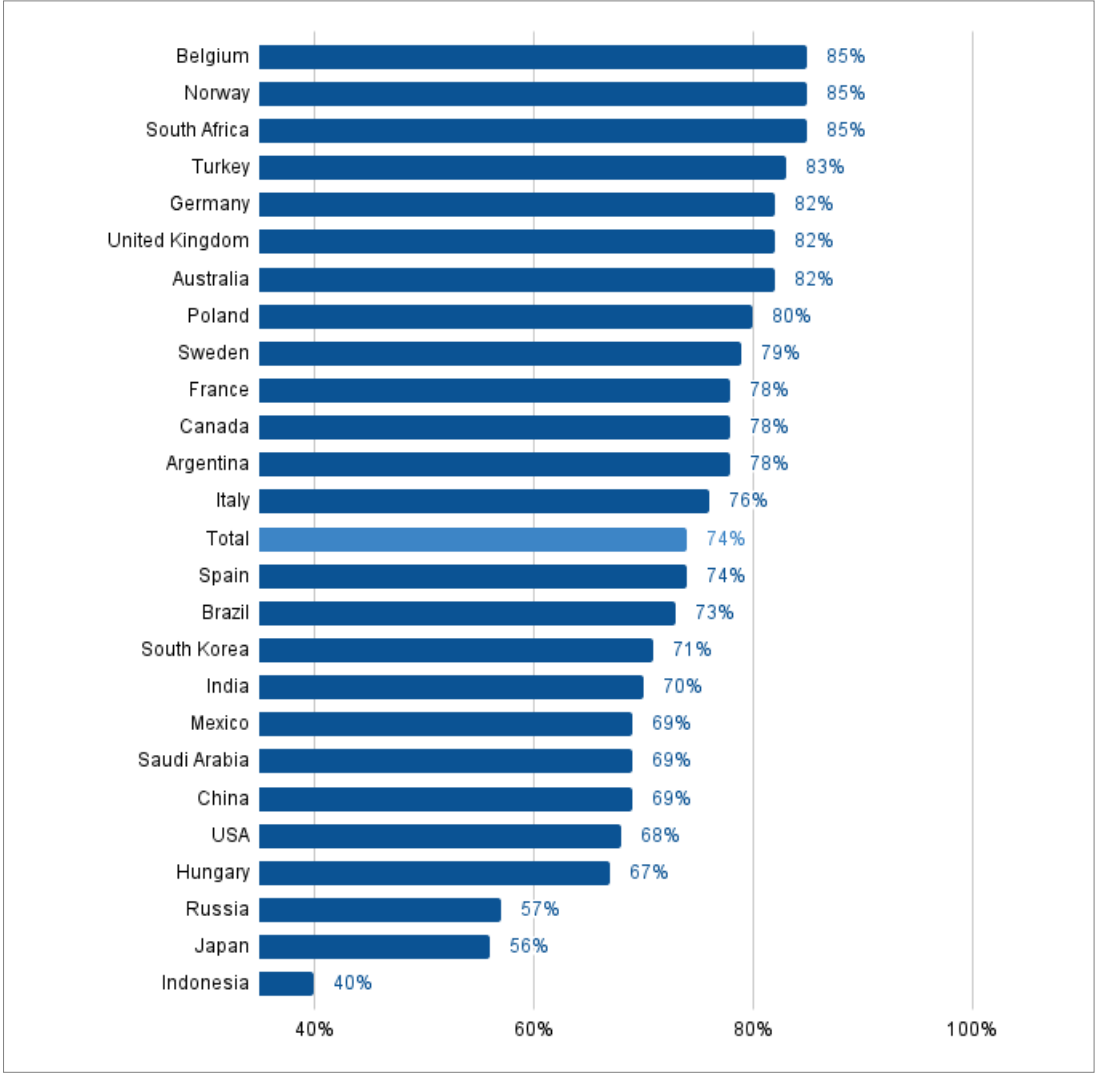

Figure 1. Search engine website usage worldwide as of August 2013, by country. Source: Based on data from Statista, 2013.

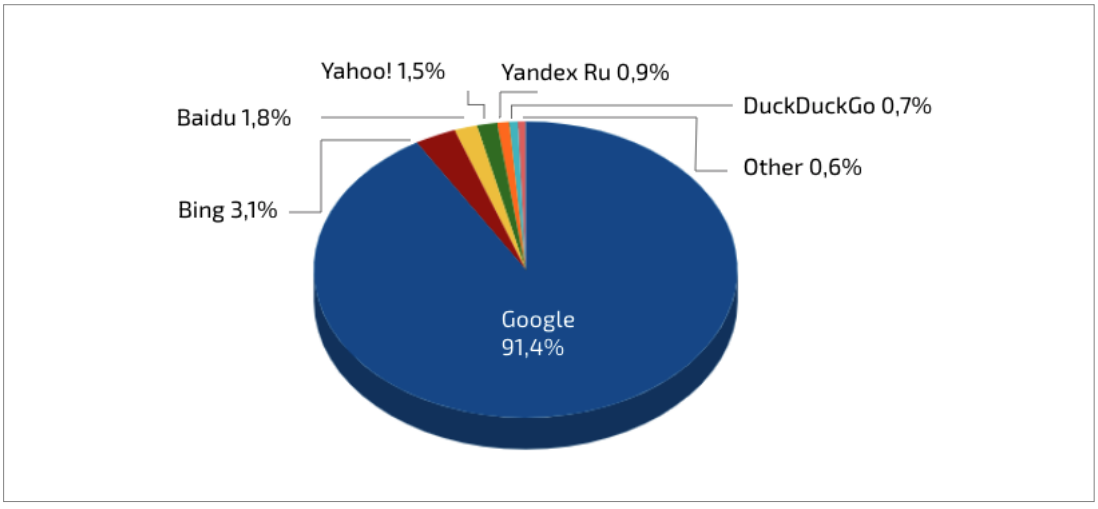

Figure 2. Search Engine Market Share Worldwide. November 2021. Source: Based on data from Statcounter 2021b

the presence of Yandex, although in May 2017 it was ousted by Google as the country's leading search engine and since then they have jousted for leadership. Indeed, in 2021, Google controlled more than 60% of the market, leaving Yandex in second position with 36% (Liveinternet, 2021). In China, the dominant search engine is by far and away Baidu, with an 87% market share in 2021, well above that of Sogou (5%) and Bing (4%), which even leave Google trailing, used by just 2% of the population (Statcounter, 2021a).

In Spain, however, Google is by far the most popular search engine, controlling more than 95% of the market, with Bing lagging a long way behind in second place with just 3% (Statcounter, 2021c).

These figures underline the importance of knowing the origin and evolution of search engines and in the case of Google – the world's market leader – the main updates made to its algorithm. Future research needs to explore further not only their characteristics but also the way they operate, specifically the ranking models used to position their results. Of special interest here is the discipline of SEO – Search Engine Optimisation – and its role in enhancing search result rankings and attracting web traffic, as well as the industry that has grown up around it, which in Spain alone comprises more than a thousand companies offering SEO services (Escandell-Poveda et al., 2021). A better understanding of this discipline will enable webmasters and web developers to understand how to increase visibility and attract audiences, essential qualities for any organization that needs to stand out in the ocean of information and resources that the Web has been converted into the decades that make up its history.

Note: This study forms part of the following doctoral thesis: Escandell-Poveda, R. (2021). *Radiografía del SEO en España: demanda laboral, oferta formativa, sector empresarial y perfil profesional del posicionamiento web.* (Radiography of SEO in Spain: labour demand, training

offer, business sector and professional profile of web positioning. [Doctoral Thesis, University of Alicante].

References

Asociación para la Investigación en Medios de Comunicación, AIMC (2020). *Navegantes en la red – encuesta AIMC a usuarios de internet.* https://www.aimc.es/otros-estudios-trabajos/navegantes-la-red/

Anklesaria, F.; Mccahill, M.; Lindner, P.; Johnson, D.; Torrey, D. & Alberti, B. (1993). The internet gopher protocol (a distributed document search and retrieval protocol). *Network Working Group Request for Comments*: 1436, 1-16. https://dl.acm.org/doi/pdf/10.17487/RFC1436

Arbildi-Larreina, I. (2005). Posicionamiento en buscadores: una metodología práctica de optimización de sitios web. *Profesional de la Información, 14*(2), 108-124. https://doi.org/10.3145/epi.2005.mar.03

Aswani, R.; Kar, A. K.; Ilavarasan, P. V. & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from twitter and SEOClerks. *International Journal of Information Management, 38*(1), 107-116. https://doi.org/10.1016/j.ijinfomgt.2017.07.005

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval.* ACM press.

Baker, L. (Febrary 27, 2021). 20+ years of SEO: A brief history of search engine optimization. *Search Engine Journal*. https://www.searchenginejournal.com/seo-101/seo-history/

Ballari, T. L. (2003). *Motores de búsqueda en internet. Un enfoque generacional*. https://web.archive.org/web/20040829010656/http://www.unlu.edu.ar/~tyr/tyr/TYR-motor/ballari-motor.pdf

Berners-Lee, T. (1989). Information management: A proposal. *The World Wide Web Consortium* (W3C). https://www.w3.org/History/1989/proposal.html

Bezhovski, Z. (2015). The historical development of search engine optimization. *Information and Knowledge Management, 5*(15) 91-96. https://iiste.org/Journals/index.php/IKM/article/view/27821

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1), 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X

Bush, V. (1945). As we may think. *The Atlantic Monthly, 176*(1), 101-108.

Casares, J. (2008). *Guía de referencia SEO*. Ojo Internet, SL.

Cutts, M. (November10, 2009). Expect caffeine after the holidays. *Matt Cutts* [personal blog]. http://www.mattcutts.com/blog/google-caffeine-update/

Devlin, J.; Chang, M.; Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Duklan, N. & Bahuguna, H. (2012). Search engine optimization techniques for organic search: A theoretical approach. *Uttaranchal Business Review, 2*(2), 45-55.

Egri, G. & Bayrak, C. (2014). The role of search engine optimization on keeping the user on the site. *Procedia Computer Science,* (36), 335-342. https://doi.org/10.1016/j.procs.2014.09.102

Eichmann, D. (1994). The RBSE spider — balancing effective search against web load. *Computer Networks and ISDN Systems,* 27(2), 308. https://doi.org/10.1016/S0169-7552(94)90151-1

Escandell-Poveda, R. (2021). *Radiografía del SEO en España: demanda laboral, oferta formativa, sector empresarial y perfil profesional del posicionamiento web.* [Tesis Doctoral, Universidad de Alicante].

Escandell-Poveda, R.; Iglesias-García, M. & Papí-Gálvez, N. (2021). Who does SEO in Spain? A cybermetric methodology for the construction of company universes. *Profesional de la Información, 30*(4). https://doi.org/10.3145/epi.2021.jul.19

Fain, D. C. & Pedersen, J. O. (2006). Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology, 32*(2), 12-13. https://doi.org/10.1002/bult.1720320206

Fernández, R. (June 3, 2021). Los buscadores online - datos estadísticos. *Statista.* https://es.statista.com/temas/3898/los-buscadores-online

Fishkin, R. (November 13, 2019). Google in 2020: From everyone's search engine to everyone's competitor. *Sparktoro.* https://sparktoro.com/blog/google-in-2020-from-everyones-search-engine-to-everyones-competitor/

Galaxy/eiNet (2021). *Galaxy/eiNet history.* https://www.einet.net/static/history.htm

García-Marco, F. (1996). Vannevar Bush, el hipertexto y el futuro del documento. In Tramullas, J. (Ed.), *Tendencias de investigación* (pp. 185-210). Documentación.

Giomelakis, D. & Veglis, A. (2016). Investigating search engine optimization factors in media websites. *Digital Journalism, 4*(3), 379-400. https://doi.org/10.1080/21670811.2015.1046992

Gonzalez, J. (March 28, 2017). Fred update – what do all losers have in common? *Sistrix.* https://www.sistrix.com/blog/googles-fred-update-what-do-all-losers-have-in-common/

Gonzalo-Penela, C. (2015). *Posicionamiento web y dinámicas de información en motores de búsqueda: propuestas de análisis y estudio comparativo de visibilidad de contenidos digitales en el caso de procesos electorales.* [Tesis Doctoral, Universitat Pompeu Fabra]. Tesis Doctorals en Xarxa. https://hdl.handle.net/10803/292730

Goodwin, D. (June 29, 2018). Why y how to track google algorithm updates. *Search Engine Journal.* https://www.searchenginejournal.com/seo-guide/track-google-algorithm-updates

Google (n/d). *De un garaje al Googleplex.* https://about.google/our-story

Google (2021). Qué es la experiencia de las páginas en los resultados de la búsqueda de google. *Centro de la Búsqueda de Google*. https://developers.google.com/search/docs/guides/page-experience

Gray, M. (1996). Credits and background. *Massachusetts Institute of Technology, MIT*. https://www.mit.edu/people/mkgray/net/background.html

Griffiths, R. T. (2002). Chapter two: From ARPANET to world wide web. *History of the internet, internet for historians (and just about everyone else)*. http://www.universelle-automation.de/1969_Washington.pdf

Grimes, C. (June 8, 2010). Our new search index: Caffeine. *Google Official Blog.* https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

Gudivada, V. N.; Rao, D. & Paris, J. (2015). Understanding search-engine optimization. *Computer (Long Beach, Calif.), 48*(10), 43-52. https://doi.org/10.1109/MC.2015.297

Gupta, A. (January 4, 2021). Google page experience update is all set to launch in may 2021 - webmasters, hang in there! *Search Engine Watch.* https://www.searchenginewatch.com/2021/01/04/google-page-experience-update-is-all-set-to-launch-in-may-2021-webmasters-hang-in-there/

Hawkins, J. (Septembrer 21, 2016). Everything you need to know about Google's 'Possum' algorithm update. *Search Engine Land.* https://searchengineland.com/everything-need-know-googles-possum-algorithm-update-258900

Internet Advertising Bureau [IAB] Spain y PricewaterhouseCoopers (2021). *Estudio de inversión publicitaria en medios digitales. Resultados 2020.* https://iabspain.es/download/50771/

Internet Live Stats (2021). Google searches in 1 second. *Internet Live Stats*. https://www.internetlivestats.com/one-second/#google-band

Jha, P.; Goyal, S.; Kumari, T. & Gupta, N. (2014). Robots exclusion protocol. *International Journal of Emerging Science and Engineering*, *2*(5). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.2053&rep=rep1&type=pdf

Johnson, J. (2021). Global market share of search engines 2010-2021. *Statista*. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Koster, M. (1994). Aliweb - Archie-like indexing in the Web. *Computer Networks and ISDN Systems*, *27*(2), 175-182. https://doi.org/10.1016/0169-7552(94)90131-7

Lara F. J. (2014) Search Engine Optimization and Ethical Leadership Strategies. In I. Gil-Pechuán, D. Palacios-Marqués, M. Peris-Ortiz, E. Vendrell, C. Ferri-Ramírez (Eds.) *Strategies in E-Business*, (pp. 79-90) Springer. https://doi.org/10.1007/978-1-4614-8184-3_7

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. *Nature (London)*, *400*(6740), 107. https://doi.org/10.1038/21987

Liveinternet (2021). *Estadística de uso de buscadores en Rusia.* https://www.liveinternet.ru/stat/ru/searches.html

Lopezosa, C.; Codina, L. & Caldera-Serrano, J. (2018). SEO semántico: Framework ISS para la optimización de sitios intensivos en contenidos. *Cuadernos De Documentación Multimedia*, (29), 97-123. https://doi.org/10.5209/CDMU.60607

Mauldin, M. I. (1997). Lycos: Design choices in an internet search service. *IEEE Expert*, *12*(1), 8-11. https://doi.org/10.1109/64.577466

Mavridis, T. & Symeonidis, A. L. (2015). Identifying valid search engine ranking factors in a web 2.0 and web 3.0 context for building efficient SEO mechanisms. *Engineering Applications of Artificial Intelligence*, (41), 75-91. https://doi.org/10.1016/j.engappai.2015.02.002

Morato-Lara, J.; Sánchez-Cuadrado, S.; Moreno-Pelayo, V. & Moreiro-González, J. A. (2013). Evolución de los factores de posicionamiento web y adaptación de las herramientas de optimización. *Revista Española De Documentación Científica*, *36*(3), e018. https://doi.org/10.3989/redc.2013.3.956

Moz (n/d). *Google algorithm update history.* https://moz.com/google-algorithm-change

Nanda, P.; Patnaik, S. & Patnaik, S. (2019). Towards algorithmic business: A paradigm shift in digital business. In Nanda, P.; Patnaik, S. & Patnaik, S. *Digital business* (pp. 3-22). Springer International Publishing. https://doi.org/10.1007/978-3-319-93940-7_1

Nayak, P. (October 25, 2019). Understanding searches better than ever before. *Google The Keyword.* https://blog.google/products/search/search-language-understanding-bert

Nierstrasz, O. (1996). W3 catalog history. *Software Composition Group*. http://scg.unibe.ch/archive/software/w3catalog/

Page, L.; Brin, S.; Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab.* http://ilpubs.stanford.edu:8090/422

Pérez-Agüera, J. (2008). Nueva generación de motores de búsqueda basados en procesamiento. *Anuario ThinkEPI*, *2*(1), 39-40. http://recyt.fecyt.es/index.php/ThinkEPI/article/view/32036/17028

Pinkerton, B. (1994). WebCrawler: Finding what people want. *Second International WWW Conference.* http://www.thinkpink.com/bp/WebCrawler/WWW94.html

Rayward, W. B. (1994). Visions of Xanadu: Paul Otlet (1868-1944) and hypertext. *Journal of the American Society for Information Science*, *45*(4), 235-250. https://doi.org/10.1002/(SICI)1097-4571(199405)45:4%3C235::AID-ASI2%3E3.0.CO;2-Y

Rovira, C.; Guerrero-Solé, F. & Codina, L. (2018). Las citas recibidas como principal factor de posicionamiento SEO en la ordenación de resultados de Google Scholar. *Profesional de la Información*, *27*(3), 559-569. https://doi.org/10.3145/epi.2018.may.09

Salazar, I. (2005). *Las profundidades de internet: Accede a la información que los buscadores no encuentran y descubre el futuro inteligente de la red*. Trea.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

Schwartz, B. (March 24, 2016). Now we know: Here are google's top 3 search ranking factors. *Search Engine Land.* https://searchengineland.com/now-know-googles-top-three-search-ranking-factors-245882

Schwartz, B. (November 10, 2020). Google page experience update to launch may 2021 with new labels in search results. *Search Engine Land.* https://searchengineland.com/google-page-experience-update-to-launch-in-may-2021-343581

Schwartz, C. (1998). Web search engines. *Journal of the American Society for Information Science, 49*(11), 973-982. https://doi.org/10.1002/(SICI)1097-4571(1998)49:11%3C973::AID-ASI3%3E3.0.CO;2-Z

Schwartz, M. F.; Emtage, A.; Kahle, B. & Neuman, C. (1992). A comparison of internet resource discovery approaches. *Computing Systems, 5*(4), 461-493. https://scholar.colorado.edu/csci_techreports/576

Seymour, T.; Frantsvog, D. & Kumar, S. (2011). History of search engines. *International Journal of Management y Information Systems, 15*(4), 47-58. https://doi.org/10.19030/ijmis.v15i4.5799

Shahzad, A.; Jacob, D. W.; Nawi, N. M.; Mahdin, H. & Saputri, M. E. (2020). The new trend for search engine optimization, tools and techniques. *Indonesian Journal of Electrical Engineering and Computer Science, 18*(3), 1568-1583. https://doi.org/10.11591/ijeecs.v18.i3.pp1568-1583

Sherman, C. (2000). The future revisited : What's new with web search. *Online (Weston, Conn.), 24*(3), 27-34. https://search.proquest.com/docview/199926335

Sherman, C. & Price, G. (2001). *The invisible web: Uncovering information sources search engines can't see.* Information Today, Inc.

Singhal, A. (February 24, 2011). Finding more high-quality sites in search. *Google Official Blog.* https://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

Smith.J & Weiss, S. (1988). Hypertext. *Communications of the ACM, 31*(7). https://doi.org/10.1145/48511.48512

Sonnenreich, W. (1997). A history of search engines. *Wiley.* https://web.archive.org/web/20040929081754/http://www.wiley.com/legacy/compbooks/sonnenreich/history.html

Statcounter (2021a). Search engine market share China. *Statcounter.com.* https://gs.statcounter.com/search-engine-market-share/all/china/

Statcounter (2021b). *Search engine market share Worldwide. Statcounter.com.* https://gs.statcounter.com/search-engine-market-share

Statcounter (2021c). *Search engine market share Spain. Statcounter.com.* https://gs.statcounter.com/search-engine-market-share/all/spain

Statista (2013). Search engine penetration worldwide by country 2013. *Statista.com.* https://www.statista.com/statistics/282842/share-of-respondents-using-a-serach-engine-regularly-worldwide-by-country/

Statista (2021). Search engine market share worldwide. *Statista.com.* https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Statista (2013). Search engine penetration worldwide by country 2013. *Statista.com.* https://www.statista.com/statistics/282842/share-of-respondents-using-a-serach-engine-regularly-worldwide-by-country/

Statista (2021). Search engine market share worldwide. *Statista.com.* https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Trigo-Aranda, V. (2004). Historia y evolución de internet. *Acta, (33)*, 1-12. https://www.acta.es/medios/articulos/comunicacion_e_informacion/033021.pdf

Van Couvering, E. (2008). The history of the internet search engine: Navigational media and the traffic commodity. *Web search* (pp. 177–206). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75829-7_11

Vise, D. A. & Malseed, D. A. (2006). *La historia de Google: Los secretos del mayor éxito empresarial, mediático y tecnológico de nuestro tiempo*. La esfera de los libros.

Wagner, G. (1996). *Upgrading search strategy skills: Unit three: Developing search strategies for searching some of the indexes of the world wide web: A continuing professional development course for library and information service managers*. Edith Cowan University. https://ro.ecu.edu.au/ecuworks/6757/

Wall, A. (2017). History of Search Engines: From 1945 to Google Today. *Search Engine History.com*. http://www.searchenginehistory.com/

Wills, R. (2006). Google's pagerank. *The Mathematical Intelligencer*, *28*(4), 6–11. https://doi.org/10.1007/BF02984696

Zhang, S. & Cabage, N. (2017). Search engine optimization: Comparison of link building and social sharing. *The Journal of Computer Information Systems*, *57*(2), 148–159. https://doi.org/10.1080/08874417.2016.1183447